Ferdowsi University of Mashhad

Journal of Computer and Knowledge Engineering

https://cke.um.ac.ir

Information and Communication Technology Association of Iran

Enhancing Channel Selection in 5G with Decentralized Federated Multi-Agent Deep Reinforcement Learning[*]

Research Article

Taghi Shahgholi[1], Keyhan Khamforoosh[2], Amir Sheikhahmadi[3], Sadoon Azizi[4]

DOI: 10.22067/cke.2024.88900.1119

Abstract The increasing popularity of vehicular communication systems necessitates efficient and autonomous decision-making to address the challenges of vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communications. In this paper, we present a comprehensive study on channelization in Cellular Vehicle-to-Everything (C-V2X) communication and propose a novel two-layer multi-agent approach that integrates deep reinforcement learning (DRL) and federated learning (FL) to enhance the decision-making process in channel utilization.

Our approach leverages the autonomy of each vehicle, treating it as an independent agent capable of making channel selection decisions based on its local observations in its own cluster. Simultaneously, a centralized architecture coordinates nearby vehicles to optimize overall system performance. The DRL-based decision-making model considers crucial factors, such as instantaneous channel state information and historical link selections, to dynamically allocate channels and transmission power, leading to improved system efficiency.

By incorporating federated learning, we enable knowledge sharing and synchronization among the decentralized vehicular agents. This collaborative approach harnesses the collective intelligence of the network, empowering each agent to gain insights into the broader network dynamics beyond its limited observations. The results of our extensive simulations demonstrate the superiority of the proposed approach over existing methods, as it achieves higher data rates, success rates, and superior interference mitigation.

Keyword *C-V2X Optimization, Multi-Agent Learning, DRL-based Channel Access, Federated Learning Integration*

1. Introduction

The Fifth Generation (5G) network has successfully transitioned into the commercial stage and is currently being swiftly deployed worldwide. Simultaneously, the proliferation of mobile devices and interactive services has resulted in a substantial surge in data traffic and user demands. Alongside human-centric communications, the prevalence of machine-to-machine (M2M) terminals is expected to escalate significantly, nearing saturation by the year 2030. Projections indicate that the number of cellular M2M terminals will reach a staggering 100 billion in 2030, approximately 10 times the figure in 2022, including more than 900 million connected cars [1].

In recent decades, the exponential increase in the number of vehicles has given rise to a range of critical issues, including traffic safety, urban congestion, and environmental pollution. In response to these challenges, there is a growing focus on establishing a transportation ecosystem that is safe, efficient, and sustainable with utilizing technologies such as 5G, autonomous and connected vehicle.

For connected vehicle, several technical solutions have been put forward where Cellular-V2X or C-V2X stands out for its ability to provide superior coverage and quality-of-service (QoS) compared to other alternatives [2]. Furthermore, by integrating advanced technologies such as millimeter-wave communication and nonorthogonal multiple access the performance of cellular V2X can be further enhanced [3]-[5]. Ensuring real-time and reliable communication for safety-critical messages poses challenges for existing centralized resource allocation in cellular networks, mainly due to diverse quality-of-service

[*] Manuscript received: 2024 July 11, Revised, 2024 October 6, Accepted, 2024 October 21.

[1] Ph.D. Student Department of Computer Engineering, Sanandaj Branch, Islamic Azad University, Sanandaj, Iran.

[2] Corresponding Author, Assistant Professor, Department of Computer Engineering, Sanandaj Branch, Islamic Azad University, Sanandaj, Iran, **Email**:k.khamforoosh@iausdj.ac.ir

[3] Associate Professor, Department of Computer Engineering, Sanandaj Branch, Islamic Azad University, Sanandaj, Iran.

[4] Associate Professor, Department of Computer Engineering, University of Kurdistan, Sanandaj, Iran.

(QoS) requirements such as ultra-reliability and low-latency.

To address these challenges, 3GPP[5] has investigated advanced resource allocation approaches for Cellular Vehicle-to-Everything (C-V2X). These approaches involve assigning independent packet priority levels to vehicular applications based on their latency and reliability requirements. Additionally, sensing-based decentralized methods have been proposed to select resource blocks with lower interference for transmission. However, it is important to note that these approaches primarily focus on dedicated resource pools and may overlook potential interference between Vehicle-to-Infrastructure (V2I) and Vehicle-to-Vehicle (V2V) communications within shared resource pools [2],[6],[7].

Machine learning has gained attention for improving C-V2X by addressing resource allocation problems and optimizing channel utilization [8]-[11]. In this paper, we have introduced a novel decentralized approach with combining federated learning and reinforcement learning for channel selection in 5G NR C-V2X connected vehicle environments enables efficient utilization of resources, enhancing system performance in terms of latency, reliability, and spectral efficiency. Our approach addresses the challenge of selecting reliable and interference-free channels, even with a high volume of vehicles. We evaluate our approach in simulation and show that it outperforms existing approaches.

The contributions of this paper that distinguish it from past works are listed below:

- Two-layer multi-agent approach: individual vehicles as agents and clusters of vehicles as higher-level coordinators
- Deep reinforcement learning (DRL) for decision-making, considering factors like CSI, queue backlog, interference, and historical selections
- Federated learning (FL) for knowledge sharing and synchronization among decentralized vehicular agents

Advantages of the Proposed Approach are Real-time adaptation and optimization of actions, utilization of collective intelligence for more informed decisions and Efficient channel utilization, minimized interference, and enhanced performance and reliability.

The subsequent sections of the paper are organized as follows: Section II provides an in-depth exploration of the background and related work, setting the foundation for our research. In Section III, we present the system model, outlining the key components and architecture of our proposed approach. Section IV delves into the problem formulation and defining the objectives of our research. The simulation and results are detailed in Section V, where we present the outcomes of our results and evaluate the performance of our proposed approach. Finally, in Section VI, we draw conclusions based on our findings, highlighting the significance of our approach.

2. Background and Related work

This paper focuses on the application of federated learning and deep reinforcement learning for channel utilization in 5G NR C-V2X. Firstly, we will provide an overview of the 3GPP standard and the channel access mechanism in 5G NR C-V2X. Next, we will delve into the use of deep reinforcement learning for resource allocation. Additionally, we will explore a novel federated learning approach and its application in channel utilization. Given the significant interest in machine learning and its application across various technologies, we conducted an extensive background and related work research to identify existing gaps in the field. By thoroughly examining the literature, we aimed to gain a comprehensive understanding of the current state-of-the-art and identify areas where further research and contributions are needed.

In Release 12, 3GPP introduced direct Device-to-Device (D2D) communications for proximity services (ProSe) using cellular technologies [12]. LTE V2X, based on the LTE air interface, was developed under Release 14 (Rel. 14) and further enhanced in Release 15 (Rel. 15). The 5G NR (New Radio) air interface served as the foundation for the development of a new cellular V2X standard under Release 16 (Rel. 16) [13].

The 5G NR standard, developed under Rel. 15, did not include sidelink (SL) aspects, which refer to direct communication between terminal nodes or User Equipment (UEs) without involving the network. However, Rel. 16 introduced V2X communications, including SL communications, based on the 5G NR air interface. This marked the availability of the first 5G NR V2X standard, focusing on connected and automated driving use cases. The goal of NR V2X SL is to support enhanced V2X (eV2X) use cases that have specific requirements not fulfilled by the LTE V2X standard.

In Release 12, two modes were defined for UE (User Equipment) transmission scheduling in V2X communications: Mode 1 and Mode 2. In Mode 1, when the UE is within the coverage of the eNB (evolved NodeB), centralized scheduling occurs at the eNB. On the other hand, in Mode 2, for D2D communication scheduling, the UE selects a radio resource from a pool configured by the cellular network or pre-configured in the UE itself to use the PC5 interface for direct communication.

Both Mode 1 and Mode 2 have a similar resource allocation structure. The data transmission is scheduled in a period called the Sidelink control period, which consists of two sets of sub-frames: Physical Sidelink Shared Channel (PSSCH) and Physical Sidelink Control Channel (PSCCH). The PSCCH is always transmitted before the PSSCH transmission to inform the receiver about the occupation of the PSSCH radio resources. This information is included in a PSCCH scheduling assignment called Sidelink Control Information (SCI). These mechanisms were designed considering the battery life of mobile devices.

However, for connected vehicle communications, different requirements need to be considered, such as latency, which D2D ProSe (Proximity Services) could not meet. Therefore, in Release 14, 3GPP introduced two new modes, Mode 3 and Mode 4, for C-V2X (Cellular Vehicle-to-Everything) to improve D2D ProSe performance.

[5] the 3rd Generation Partnership Project

In Mode 3, similar to Mode 1, provides centralized scheduling, ensuring efficient resource utilization, but it requires vehicles to be within network coverage and introduces cellular uplink and downlink signaling overhead. Mode 4, on the other hand, enables vehicles to operate outside network coverage and make independent sub-channel selections using the sensing based SPS scheme where vehicles utilize sensing techniques to identify and select suitable sub-channels for transmission [16].

However, the PSCCH and PSSCH allocation in Mode 3 and 4 are completely different from Mode 1 and 2. In Mode 3 and 4, the resources are divided into sub-channels, and the first resource blocks are PSCCH pools, while the rest of the resource blocks are PSSCH for data transmission (Transport Blocks).

In Mode 4, UEs can select their sub-channels using a new mechanism called sensing-based semi-persistent scheduling, which significantly improves the estimation of available sub-channels.

For the access mechanism in C-V2X, it supports 10 and 20 MHz channels with Single Carrier Frequency Division Multiple Access (SC-FDMA). The channels are divided into Resource Blocks (RBs), sub-channels, and sub-frames. Resource blocks are 180 kHz wide in frequency and consist of 12 sub-carriers of 15 kHz. The sub-frames are defined as 1ms long, and a sub-channel is a group of resource blocks in the same sub-frame. In C-V2X, there are two sub-channelization schemes: Adjacent PSCCH + PSSCH and Nonadjacent PSCCH + PSSCH. In the adjacent scheme, the Sidelink Control Information (SCI) and its associated Transport Block (TB) are in adjacent resource blocks. The first two resource blocks of the first sub-channel are used for the SCI, while the transport block occupies several sub-channels in the next resource blocks. In the nonadjacent scheme, resource blocks are divided into pools, with a dedicated pool for transmission of the SCIs and other pools used for TBs transmissions. In Sidelink communications, each vehicle selects a transmission resource block without communicating with the Base Station (BS) and directly sends data to other vehicles.

During the SPS process, a vehicle first senses the transmissions in its vicinity to assess the availability and quality of different sub-channels. Based on this information, the vehicle identifies candidate resources within a designated Selection Window (SW). The SW includes a range of subframes where the vehicle can find sub-channels that can accommodate its transmission. Once the candidate resources are identified, the vehicle excludes specific resources based on the sensed interference or other criteria. The remaining sub-channels within the SW are then considered for transmission. The vehicle reserves these selected sub-channels for its subsequent transmissions using the Resource Reservation Interval (RRI) included in the Sidelink Control Information (SCI). This approach allows for efficient utilization of available sub-channels and helps mitigate interference in V2X Sidelink communications [17].

While SPS is a simple and effective method, it has several limitations including lower QoS in higher density, faces challenges in handling packet collisions due to imprecise sensing results caused by the hidden-terminal problem, the half-duplex constraint preventing the detection of other vehicles using the same resources, the increased likelihood of collisions in high-density scenarios, etc.

In this paper, we propose a novel approach to improve the resource allocation process in C-V2X Mode 4, where we integrate the clustering technique, federated learning, and multi-agent deep reinforcement learning algorithm into SPS, enabling vehicles to intelligently select radio resources and avoid resource conflicts. This approach leverages federated learning and clustering techniques to enhance resource allocation performance in C-V2X communications.

To overcome the SPS limitations, in recent years, novel resource allocation schemes in C-V2X communications has been explored. These schemes employ both centralized and decentralized approaches, with a primary focus on optimizing parameter configurations, enhancing the resource sensing process, and improving the resource allocation process.

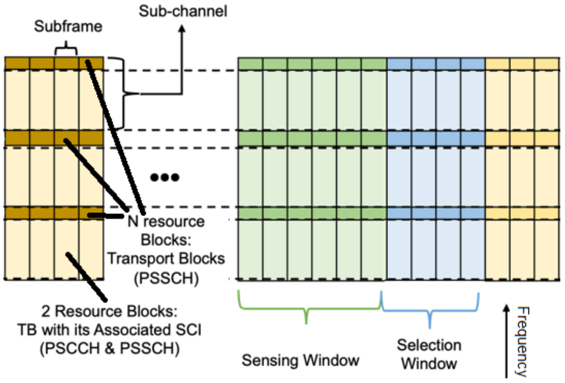
Sub-channel
Subframe
N resource Blocks: Transport Blocks (PSSCH)
2 Resource Blocks: TB with its Associated SCI (PSCCH & PSSCH)
Sensing Window
Selection Window
Frequency

Figure 1. summarizes the channel resource management in C-V2X standard.

For centralized approach, in [15]- [18], a power control algorithm was developed based on spatiotemporal traffic patterns. This algorithm aims to satisfy the delay and reliability requirements of V2V services while reducing the overhead of periodic Channel State Information (CSI) reports to the Base Station (BS). Graph theory has also been utilized in [19]-[21] to enhance system throughput. In the study [19], the ergodic capacity of Vehicle-to-Infrastructure (V2I) communications and the reliability of V2V communications were analyzed by considering the statistics of fast fading components. Based on these analyses, the researchers proposed centralized resource allocation and power control algorithms to meet diverse QoS requirements.

In [22], the authors addressed the challenge of channel uncertainty caused by delayed Channel State Information (CSI) feedback to the Base Station (BS). They analyzed the correlation of fast-changing channels and proposed a joint channel allocation and power control algorithm. The objective was to maximize system throughput while meeting the delay and reliability requirements of each V2V link. However, these existing algorithms rely on V2V links reporting their local information, which can result in significant signaling overhead as the number of vehicles increases. Moreover, resource allocation has been formulated as combinatorial optimization problems with nonlinear constraints, posing challenges for traditional optimization methods.

To overcome the centralized problems, recently, several studied proposed decentralized approaches. The authors in [23] introduced a decentralized resource allocation approach for vehicle-to-vehicle communications. In this approach, each V2V transmitter acted as an independent agent and made decisions autonomously based on local observations. The system was able to optimize resource allocation in a decentralized manner, considering the unique characteristics and requirements of each V2V transmitter.

In [24], the authors propose a novel semi-distributed transmission paradigm for NR V2X to achieve high reliability. It includes distributed clustering, autonomous inter-cluster resource selection, and centralized intra-cluster communication. In [25], the authors focused on reducing transmission failures and designed a distributed network coding-based medium access control protocol (NC-MAC) for reliable V2V beacon broadcasting. By combining preamble-based feedback, retransmissions, and network coding, the NC-MAC protocol enhances broadcasting reliability.

These studied have primarily focused on short-term optimization of resource allocation, overlooking the potential long-term performance gains. It is essential to consider a broader perspective and incorporate long-term strategies into resource allocation algorithms to achieve sustained performance improvement over time. By considering factors such as network dynamics, future traffic patterns, and system scalability, resource allocation approaches can be designed to optimize not only immediate resource allocation decisions but also their impact on overall network performance in the long run.

To address this challenge, researchers have turned to machine learning techniques, especially deep reinforcement learning (DRL) and more recently federated learning. These approaches offer effective solutions for tackling sequential decision-making problems. By combining deep learning and reinforcement learning, DRL algorithms enable the learning of optimal strategies in complex environments with long-term consequences for actions. Federated learning, on the other hand, allows distributed devices to collaboratively learn from their local data without sharing it centrally. These powerful techniques hold promise for addressing various challenges and optimizing decision-making processes in complex environments such as connected vehicles.

Deep Reinforcement Learning have gained significant traction in the field of wireless communications as they provide effective solutions to the challenges encountered by traditional optimization methods specially for resource allocation [26].

In [27], the authors introduced the C-Decision architecture for resource allocation in V2X networks. It combines centralized decision making and distributed resource sharing to maximize the sum rate. Vehicles compress their information using deep neural networks and send it to the centralized decision unit. The decision unit employs a deep Q-network for resource allocation and balances V2V and V2I links. To overcome the problem of high collision probability in conventional SPS with a fixed reservation process during high traffic density, [28] proposes a Q-learning based SPS (Q-SPS) algorithm. Q-SPS intelligently adjusts the reservation probability using reward feedback, adapting to the dynamic C-V2X network environment. However, this approach deviates from the fundamental assumption of RL, which requires a stationary environment. In this case, a single vehicle is unable to update the evolving policies of other vehicles, thereby compromising the effectiveness of the approach [29][30].

In [31], the authors employed the DRL algorithm as a means to allocate resource blocks (RBs) and minimize signal collisions during transmissions. However, this approach did not consider the heterogeneous nature of quality-of-service (QoS) requirements across different types of messages where various message types may have distinct QoS demands, such as latency, reliability, and priority. Ignoring these varying requirements could lead to suboptimal resource allocation decisions and potential degradation of overall network performance.

Several studies have focused on addressing these challenges through the implementation of multi-agent DRL techniques. In [32], authors address resource allocation challenges in V2X communications and proposes two algorithms. The first algorithm uses deep reinforcement learning (DRL) with deep Q-network (DQN) and deep deterministic policy-gradient (DDPG) to improve performance for V2I and V2V links. The second algorithm, based on meta-learning, enhances adaptability to dynamic environments. [33] focuses on spectrum allocation in V2X networks using a graph representation. A graph neural network (GNN) is employed to extract low-dimensional features from the graph. Multi-agent RL is then used to allocate spectrum based on the learned features and deep Q-network is utilized for optimizing the sum capacity of the V2X network. Several other studies have employed different types of DRL algorithms and proposing scheme utilizing multiagent deep deterministic policy gradient, proximal policy optimization (PPO)-based multi-agent reinforcement, deep deterministic policy-gradient (DDPG), long short-term memory (LSTM) etc. to optimize the allocation of resources in V2X communications [34]-[39].

While DRL approaches have shown promise for resource allocation in vehicular networks, there are challenges to consider regarding training efficiency, particularly in highly dynamic and large-scale environments. The complexity and rapid changes in network conditions pose difficulties in achieving fast and accurate convergence during the training process. Addressing these issues is crucial to ensure the practicality and scalability of DRL-based resource allocation algorithms in real-world vehicular communication scenarios.

Moreover, using a fully distributed DRL method can lead to convergence at local optima, while fully centralized DRL methods are not suitable for vehicular networks due to the significant delay caused by information exchange with central nodes, particularly for delay-sensitive applications. Furthermore, the computational complexity of centralized DRL increases substantially with a larger number of vehicles. Hence, a more viable approach is to combine the strengths of centralized and distributed DRL algorithms to effectively support direct communications in vehicular networks.

To address the limitations of traditional DRL approaches, researchers have recently turned to the emerging technique of federated learning [40] for resource allocation problems. Federated learning enables collaborative learning across multiple decentralized devices or nodes without the need to share raw data. The application of federated learning in resource allocation holds great promise in improving the performance and adaptability of wireless networks, as it leverages the collective intelligence of distributed devices to optimize resource allocation decisions.

The process of federated learning consists of three steps: First, the FL server determines the training task and distributes the initial global model to selected distributed devices. These devices then utilize their local data to train their individual models, aiming to minimize the loss function based on the initial global model. After several rounds of local training, the devices upload their local models to the FL server. Finally, the FL server aggregates these local models and sends back the updated model to the data owner. This process is repeated until the global loss function converges or a desired training accuracy is achieved. By conducting local model training on decentralized devices using their own raw data and infrequent model aggregation at the centralized server, federated learning significantly enhances the performance of model training [41]-[43].

Federated learning is a promising approach that has gained attention in various domains. However, its application in C-V2X resource allocation is still relatively new, and there are limited studies exploring its potential in this context. Authors in [44] introduce federated learning into a MEC-assisted vehicular network framework with focuses on participant selection, computing resource allocation optimization, and a distributed computing resource allocation method.

Authors in [45] propose a deep reinforcement learning (DRL)-based federated learning (FL) approach for decentralized resource allocation in an underlay mode D2D-enabled wireless network. The aim is to maximize sum capacity, minimize power consumption, and ensure quality of service (QoS) for both cellular and D2D users. Furthermore, a joint optimization problem involving transmission mode selection and resource allocation is investigated in [2], and formulated as a Markov decision process. The authors also proposed a DRL-based decentralized algorithm to maximize the sum capacity of V2I users while meeting latency and reliability requirements for V2V pairs. Also, to overcome training limitations, a two-timescale federated DRL algorithm is introduced, utilizing a graph theory-based vehicle clustering algorithm on a large timescale and federated learning on a small timescale.

None of the mentioned studied have not taken into account the scenario of shared spectrum in V2V communication, where the cellular channels are already assigned, and the traffic is highly congested. Authors in [46] proposed approach is a federated multi-agent deep reinforcement learning (FedMARL) method that optimizes channel selection and power control for V2V communication. By leveraging both deep reinforcement learning (DRL) and federated learning (FL), the approach ensures reliability, delay requirements, and maximizes cellular link transmit rates. Individual V2V agents are constructed using the dueling double deep Q-network (D3QN) and trained collaboratively with a designed reward function. Federated learning is incorporated to address training instability in the multi-agent environment. An important limitation of this study is that it relied on static and pre-defined resources for decentralized channel access. The use of fixed resources may not effectively adapt to dynamic channel conditions or varying traffic demands in real-time.

3. System Model

A. Network model

The system model we consider is a network consisting of various clusters of vehicles, including Vehicle-to-Vehicle (V2V), Vehicle-to-Infrastructure (V2I), Roadside Units (RSUs), and a Base Station (BS) as shown in Fig. 2. Each cluster is denoted as $C = c_1, c_2, \dots, c$⍰, with k representing the total number of clusters in the network. Within each cluster C□, there is a set of vehicles denoted as V⍰ $=$ $v_1, v_2, \dots v_{nk}$ where n□ is the total number of vehicles in cluster C□. The set of RSUs is denoted as $\mathcal{J} = j_1, j_2, \dots,$ j_m with m representing the total number of infrastructure units.

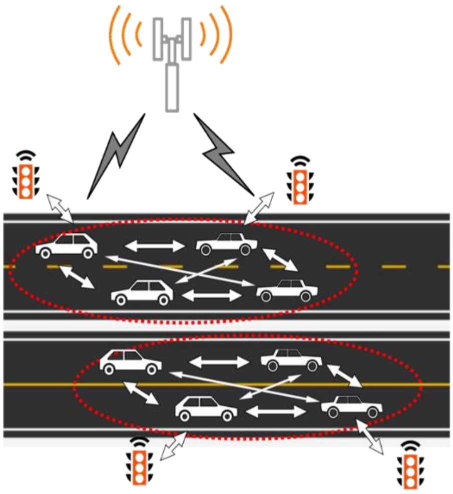

Figure 1. High-level System Model

In this network model, the time is divided into slots, indexed by $t = 1, 2, \dots$, where each slot has a duration of τ. This slotted communication system provides a structured framework for organizing and managing the transmission and reception of data in the system. In our model, each base station serves $\mathrm{T_c}$ number of cellular users, which include vehicles and RSUs. The selected users within each cluster communicate with the base station using allocated cellular channels. However, within the clusters, V2V and V2I communication take place by reusing the allocated channels. This enables efficient utilization of resources and facilitates direct communication between vehicles and infrastructure units within the same cluster.

In our network model, we consider the channel power gains between different links and the interference could cause when they use same channel. Specifically, the channel power gain from the transmitter of the $\alpha - th$ V2I link to the BS over the $f - th$ RB outside the cluster is denoted as $g_{\alpha,B}[f]$.

We can define the SINR of the $\alpha - th$ V2I link at the BS, and $\beta - th$ V2V link at the $f - th$ RB and in time slot i as [19]:

$$\mathrm{SINR}_{i,\alpha,f} = \frac{P_{i,\alpha,f} \times \ g_{\alpha,B}[f]}{\sigma^2 + \sum_j (\rho_{j,\beta,f} \times P_{j,\beta,f} \times g_{\beta,B}[f])} \qquad (1)$$

where the $\rho_{j,\beta}$ is is the spectrum allocation indicator where $\rho_{\beta,f} \ = 1$ means the $\alpha - th$ V2I links is transmitting over the $f - th$ RB and $\rho_{j,\beta} \ = 0$ means it does not transmit

Similarly, for $\beta - th$ V2V link at the V2V receiver over the $f - th$ RB we can define:

$$SINR_{j,\beta,f} = \frac{P_{j,\beta,f} \times g_{\beta}[f]}{\sigma^2 + \sum_j (\rho_{j,\beta,f} \times P_{j,\beta,f} \times g_{j,\beta}[f]) + \sum_\gamma (\rho_{\gamma,\beta} \times P_{\gamma,f} \times g_{\gamma,B}[f])} \qquad (2)$$

In these equations, $P_{i,\alpha,f}$ and $P_{j,\beta,f}$ represent the transmit powers of the $\alpha - th$ V2I and the $\beta - th$ V2V transmitter over the $f - th$ RB, respectively. The indicator $\rho_{c,m,f} \in \{0,1\}$ is the spectrum allocation indicator, where $\rho_{i,\alpha,f} = 1$ implies that the $\alpha - th$ V2I link is transmitting over the $f - th$ RB, and $\rho_{i,\alpha,f} = 0$ otherwise. Similarly, the spectrum allocation indicator for the $\beta - th$ V2V link, $\rho_{j,\beta,f}$ is defined in a similar manner. In addition, the $P_{\gamma,f}$ represent the interfering channel from the $\gamma - th$ V2V transmitter to the $\beta - th$ V2V receiver over the $f - th$ RB.

B. QoS Requirements

- ***Delay for V2V Pairs***

In our system, each transmitter of a V2V link is equipped with a finite-length buffer, and safety-related packets are generated at a constant rate λ (bits/s). However, due to varying transmit rates $R_{tk} = W \log_2(1 + \gamma_{tk})$ at different time slots, there can be a mismatch between packet generation and instantaneous throughput. Consequently, queues can build up at the transmitters of V2V pairs, resulting in increased queuing delays.

At the beginning of time slot t, the queue length of the $\beta - th$ V2V pair, denoted by $Q_{t\beta}$, is determined by the following equation [46]:

$$Q_{t\beta} = \max(0, Q_{t-1,\beta} + \tau\lambda - \tau R_{t-1,\beta}) \qquad (3)$$

where $Q_{t-1,\beta}$ is the queue length at the transmitter in the previous time slot $(t - 1)$, $\tau\lambda$ represents the number of bits arrived at the queue per slot, and $\tau R_{t-1,\beta}$ denotes the number of bits sent to the corresponding receiver in the previous time slot $(t - 1)$.

We focus on the queuing delay as it dominates the delay over a V2V link. Based on Little's Law, the average queuing delay is proportional to the queue length. Let D_{max} present the tolerable transmission delay for V2V packets, and the number of packets experiencing delays longer than D_{max} given by $Q_{max} = \lambda D_{max}$. Therefore, the delay constraint for the $\beta - th$ V2V pair can be rewritten to ensure a steady-state queue length with a tolerable probability threshold [46]:

$$\Pr(D_{t\beta} \geq D_{max}) \leq \Pr(Q_{t\beta} \geq Q_{max}) \leq p_o \qquad (4)$$

p_o is the probability threshold.

By applying Markov's inequality, which states that $\Pr(X \geq a) \leq \frac{E[X]}{a}$ for a non-negative random variable $(X$ and $(a > 0$, we can further strengthen the constraint on the queue length (4) in the following manner [46]:

$$\overline{Q}_k = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} E|Q_{t\beta}| \leq p_o \cdot Q_{max} \qquad (5)$$

here $\overline{Q}_k$ represents the time-averaged queue length of the transmitter. The upper bound on the time-averaged queue length ensures that V2V packets can be delivered within the specified time constraints.

- *Reliability Requirements*

In order to ensure reliable transmission of V2V pairs, the SINR outage probability is considered as a key metric. The outage probability represents the likelihood of the instantaneous SINR falling below a specified threshold γ_o, indicating a loss of signal quality due to wireless channel fading. Evaluating the reliability of V2V transmissions, the outage probability is compared against a predetermined threshold value. A V2V link γ is considered reliable if its outage probability is below a specified threshold p_0. Mathematically, this condition can be expressed as:

$$\Pr\{\gamma_{\beta,f} \leq \gamma_o\} \leq p_0 \qquad (6)$$

To achieve reliable transmission, vehicles have the option to adjust their power levels or switch to less congested channels if their outage probability constraint cannot be met. By incorporating the SINR of V2V/V2I pairs, the constraint for outage probability can be expressed as:

$$\Pr\left\{P_{j,\beta,f} \times \ g_{\beta}[f] \leq \sum_{\gamma} (\rho_{\gamma,\beta} \times P_{\gamma,f} \times g_{\gamma,B}[f]) + \gamma_0((\rho_{j,\alpha,f} \times P_{i,\alpha,f} \times \ g_{\alpha,B}[f])\right\} \leq p_o \qquad (7)$$

If $x_1, x_2, \dots, x_n$ are independent exponentially distributed random variables with expected values $E[x_i] = \frac{1}{\lambda_i}$, where $i = 1, 2, \dots, n$, then the probability that x_1 is less than or equal to the sum of $x_2, x_3, \dots, x_n$ plus a positive constant c can be expressed as [47].

$$\Pr\{x_1 \leq x_2 + x_3 + x_4 + \ldots + x_n + c\} = 1 - e^{-\lambda_1 c} \prod_{i=2}^{n} \frac{1}{1 + \frac{\lambda_1}{\lambda_i}} \tag{8}$$

Considering that the fast fading varies independently between slots, it follows an exponential distribution. Thus, $P_{j,\beta,f} \times g_\beta[f]$, and $\gamma_0 \times (\rho_{\gamma,\beta} \times P_{\gamma,f} \times g_{\gamma,B}[f])$, also follow exponential distribution and can be express as:

$$E\left[P_{j,\beta,f} \times g_\beta[f]\right] = \frac{1}{\lambda_1} \tag{9}$$

and

$$E[\gamma_0 \times (\rho_{\gamma,\beta} \times P_{\gamma,f} \times g_{\gamma,B}[f])] = \frac{1}{\lambda_\gamma} \quad \gamma \neq 1 \tag{10}$$

Based on equation (8), we can derive the expression for the outage probability as follows:

$$1 - \exp\left(-\frac{\gamma_0 \times \sigma^2}{P_{\mathrm{j,\beta,f}} \times \zeta}\right)\left(1 + \frac{P_{j,\beta,f} \times \gamma_0}{P_{\gamma,B} \times \zeta}\right)\prod_{k=k}^{K}\left(1 + \frac{P_{j,\beta,f} \times \gamma_0}{P_{j,\beta,f} \times \zeta}\right) \leq p_0 \tag{11}$$

ere ζ is the frequency-independent large-scale fading effect, which encompasses path loss and shadowing, can be characterized by the path loss model defined as $128.1 + 37.6\ log_{10} d$, as specified in 3GPP TR 36.885 [48]. Here, d represents the distance between the transmitter and receiver.

Considering the inequality [48]:

$$e^k \times \prod_{i}^{n} x_i \leq e^{k+(x_1+x_2+\cdots+x_n)} \tag{12}$$

upper bound of (11) can be defined and reliability constraint can be expressed as follows [46]:

$$\frac{P_{j,\beta,f} \times g_\beta[f]}{\sigma^2 + \sum_j(\rho_{j,\beta,f} \times P_{j,\beta,f} \times g_{j,\beta}[f]) + \sum_\gamma(\rho_{\gamma,\beta} \times P_{\gamma,f} \times g_{\gamma,B}[f])} \geq \frac{\gamma_O}{\ln\left(\frac{1}{1-p_O}\right)} \quad (13),$$

I. Problem Formulation

In order to satisfy the varying QoS demands of different vehicular links, such as high capacity for V2I connections and reliable performance for V2V connections, we aim to maximize the total capacity of the V2I links while ensuring a reliability level of reliability for each V2V link. This leads us to formulate the spectrum and power allocation problems as follows:

The objective is to maximize the expression:

$$\max_{\rho_{i,\alpha,f}, \rho_{j,\beta,f}} \frac{1}{T} \sum_{m,f} \rho_{i,\alpha,f} \log_2(1 + \gamma_{i,\alpha,f}) \tag{14}$$

subject to the following constraints:

$$\rho_{j,\beta,f} Pr(\gamma_{j,\beta,f} \leq \gamma_0) \leq p_0, \forall k, f \tag{14-a}$$

$$\sum_{\beta} \rho_{j,\beta,f} = 1, \forall f \tag{14-b}$$

$$\sum_{f} \rho_{j,\beta,f} = 1, \forall \beta \tag{14-c}$$

$$Q_t^\beta \leq Q_{max}, \forall \beta \tag{14-d}$$

$$\sum_{f} \rho_{j,\beta,f} = 1, \forall \beta \tag{14-e}$$

$$\sum_{f} \rho_{i,\alpha,f} P_{i,\alpha,f} \leq P_{i,max}, \forall \alpha \tag{14-f}$$

$$\sum_{f} \rho_{j,\beta,f} P_{j,\beta,f} \leq P_{f,max}, \forall \beta \tag{14-g}$$

$$P_{i,\alpha,f} \geq 0, P_{j,\beta,f} \geq 0, \forall \alpha, \beta, f \tag{14-h}$$

$$\rho_{i,\alpha,f}, \rho_{j,\beta,f} \in 0,1, \forall \alpha, \beta, f \tag{14-i}$$

This optimization problem aims to maximize the sum capacity of the V2I links while satisfying the reliability constraint for V2V links, along with various spectrum and power allocation constraints. Problem (14) represents a complex optimization problem that is challenging both mathematically and computationally. It involves making joint decisions on channel selection and power allocation over time, which requires considering various combinations and scenarios. Traditional centralized approaches struggle to handle this problem effectively, primarily due to the difficulty in acquiring accurate and up-to-date channel state information (CSI) for all links in real-time.

To address these challenges, we propose a federated-based decentralized solution leveraging the power of DRL. By applying DRL techniques, we transform the original problem into a multi-agent framework, where each V2V pair acts as an independent agent responsible for its own resource allocation strategy.

In this decentralized approach, the communication pairs autonomously make decisions on channel selection and power allocation based on local observations and rewards obtained through interactions with the environment. Through continuous learning and policy updates, each agent improves its resource allocation strategy over time, optimizing the overall system performance.

By distributing the decision-making process among the individual communication links, the decentralized DRL approach offers several advantages. It reduces the computational burden by distributing the optimization task across multiple agents. It also mitigates the need for centralized coordination and real-time CSI exchange, which can be challenging in practical scenarios.

Furthermore, the decentralized nature of the DRL approach enables scalability and adaptability to dynamic network conditions. Each agent can quickly adapt to changes in the environment and adjust its resource allocation strategy, accordingly, ensuring efficient and reliable communication.

A. DRL Formula

In order to understand the application of DRL in our context, let's introduce the fundamental concepts of DRL and its extension to a multi-agent setting. DRL involves training an intelligent agent to make optimal sequential decisions by interacting with its environment through trial and error.

At each time step *t*, the agent observes its surrounding environment and receives an observation s^t ∈ S. Based on this observation, the agent selects an action at $a^t \in A$ according to a policy π: S ⟶ A, which determines the probability of taking a specific action given a certain state. The environment is influenced by the executed action, leading to a transition to the next state $s^{t+1} \in S$. In response to the action, a reward $r^t = R(s^t, a^t)$ is provided to the agent, evaluating the impact of the chosen action and enabling the agent to adjust its policy accordingly. Each interaction between the agent and the environment creates an experience, represented by a tuple (s^t, a^t, r^t, s^{t+1}).

By accumulating these experiences and employing appropriate DRL algorithms, the agent can learn to make informed decisions over time. Through a series of trials and adjustments, the agent improves its policy, maximizing the cumulative rewards obtained from the environment. In a multi-agent setting, each agent follows this learning process independently, interacting with its own observations, actions, and rewards.

Q-learning is a widely used DRL algorithm that aims to maximize the expected cumulative reward, also known as the Q-value, based on a given policy π. The Q-value denoted as $Q_\pi(s, a)$, represents the expected total reward an agent can achieve by taking action a in state s and following policy π thereafter. Mathematically, the Q-value can be defined as the expected sum of discounted future rewards, as shown in Equation (15).

$$Q\pi(s,a) \\ = E_\pi[\textstyle\sum_{k=0}^{\infty} \beta^k r_{t+k+1} \mid s_t = s, a_t = a\,] \tag{15}$$

where E_π denotes the expectation under policy π, r_t is the reward obtained at time step t, and β is a discount factor that determines the importance of future rewards.

Q-learning is a popular DRL algorithm that maximizes the expected cumulative reward (Q-value) based on a given policy π. The Q-value $Q^\pi(s, a)$ represents the expected cumulative reward when taking action a in state s following policy π. The goal of Q-learning is to find the optimal policy that maximizes the Q-value for each state-action pair.

In Q-learning, an agent maintains a Q-table to store the Q-values for all possible state-action pairs. The Q-values are updated iteratively based on the observed rewards and the agent's learning rate. At each step, the agent selects an action a based on the current state s and updates the Q-value using the following equation:

$$Q(s,a) \leftarrow (1-\alpha) \cdot Q(s,a) + \alpha \\ \cdot \left(r + \gamma \cdot \max_{a'} Q\,(s', a')\right) \tag{16}$$

where α is the learning rate, *r* is the immediate reward obtained by taking action a in state s, γ is the discount factor, s' is the next state, and a' is the next action. This update equation combines the current Q-value with the discounted future Q-value of the next state-action pair, scaled by the learning rate.

To handle large-scale problems with high-dimensional state and action spaces, Deep Q-Network (DQN) was introduced. Instead of using a Q-table, DQN employs a deep neural network to approximate the Q-value function. The neural network takes the state s as input and outputs the Q-values for all possible actions. The parameters of the neural network are updated through gradient descent using a loss function that minimizes the difference between the predicted Q-values and the target Q-values. In the multi-agent setting, each agent interacts with the environment and learns independently. However, their actions collectively influence the environment's dynamics. To address this, the concept of multi-agent reinforcement learning (MARL) is introduced. MARL allows agents to learn and adapt their policies by considering the joint actions and observations of other agents.

One approach in MARL is Independent Q-Learning (IQL), where each agent maintains its own Q-values and learns independently. The Q-values are updated based on the observed rewards and the Q-values of other agents' actions. The update equation for agent i can be written as:

$$Q_i(s, a_i) \leftarrow (1-\alpha_i) \cdot Q_i(s, a_i) + \alpha_i \cdot \\ \left(r_i + \gamma \cdot \max_{a_i'} Q_i\,(s', a_i')\right) \tag{17}$$

where a_i is the action taken by agent i, r_i is the immediate reward obtained by agent i, and s' is the next state. The Q-value update is similar to the single-agent case, but it takes into account only the individual agent's actions and rewards. It is important to note that MARL introduces additional challenges such as coordination among agents and balancing exploration and exploitation. Various algorithms and techniques have been proposed to address these challenges, including centralized training with decentralized execution, communication among agents, and opponent modeling.

In our paper, we propose a decentralized DRL approach that operates within a collaborative reward setting. The aim of our research is to enable multiple agents to learn and make optimal decisions in a distributed manner, while striving to maximize the global cumulative reward. By leveraging DRL techniques, we address the challenges associated with large-scale problems and the complex interactions between agents.

The key idea behind our approach is to train individual agents to independently learn their own policies based on local observations of the environment. Each agent selects actions based on its own policy, contributing to a joint action that impacts the overall state transition. The agents receive a common reward signal, encouraging them to coordinate their actions towards achieving a collective objective meeting the QoS requirements. The learning model consists of the following key elements:

1) **State:** The global state, denoted as S_t, captures the channel conditions of all V2V / V2I pairs, as well as the resource allocation actions and reusing the channel of the V2V pairs. Each V2V agent (or agent k) has access to a local state, which includes the channel coefficient of the V2V pair $\left(h_{t,k}[m]\right)$, the channel coefficients of the V2I $\left(h_{t,m,B}[m]\right)$, the channel selection of neighboring V2V pairs in the previous slot $\left(N_{m,t-1,k}\right)$, and the current queue length at the transmitter $\left(Q_{t,k}\right)$. The state space size per V2V pair is $3M + 1$, where M represents the number of available channels.
2) **Action:** Each V2V agent takes actions that determine the channel selection $\zeta_{k,m}$ and transmit power P_k.The transmit power is discretized into $N_p + 1$ levels, and the action space dimension is $M \times \left(N_p + 1\right)$.
3) **Reward:** The reward function, denoted as R_t, is designed to maximize the total capacity of the V2I links while ensuring a reliability level of reliability for each V2V link. It is defined as:

$$R_t = \Gamma_1 \sum_{k\in K} U\left(\frac{P_{j,\beta,f} \times g_{\beta}[f]}{\sigma^2 + \sum_j\left(\rho_{j,\beta,f} \times P_{j,\beta,f} \times g_{j,\beta}[f]\right) + \sum_\gamma\left(\rho_{\gamma,\beta} \times P_{\gamma,f} \times g_{\gamma,B}[f]\right)} - \frac{\gamma_o}{\ln\left(\frac{1}{1-p_o}\right)}\right) + \Gamma_2 \sum_{k\in K} U\left(Q_{t,k} - Q_{\max}\right) + \Gamma_3 \sum_{m\in M} U\left(R_{t,m} - R_{\min,m}\right) \tag{18}$$

Here, $(\Gamma_1), (\Gamma_2), and (\Gamma_3)$ are parameter coefficients that balance the importance of different components in the reward function. $R_{t,m}$ represents the achieved rate of V2I communication on channel (m), while $(R_{\min,m})$denotes the minimum required rate for V2I communication on channel (m). $(Q_{t,k})$ represents the queue length at the transmitter of V2V pair (k), and $(Q_{\max})$is the maximum tolerable queue length.

The last term of the reward function represents the reward for the V2V pairs. It incorporates the SINR

$$\left(\frac{P_{j,\beta,f} \times g_{\beta}[f]}{\sigma^2 + \sum_j\left(\rho_{j,\beta,f} \times P_{j,\beta,f} \times g_{j,\beta}[f]\right) + \sum_\gamma\left(\rho_{\gamma,\beta} \times P_{\gamma,f} \times g_{\gamma,B}[f]\right)}\right)$$

and compares it with the threshold $\left(\frac{\gamma_o}{\ln\left(\frac{1}{1-p_o}\right)}\right)$. The function $U(x)$ provides a penalty if the reward condition is not satisfied, where x can be either positive or negative.

The reward function aims to maximize the achieved rates of V2I communication, minimize the queue lengths of V2V pairs, and ensure satisfactory SINR levels for V2V links. The coefficients $(\Gamma_1), (\Gamma_2), and (\Gamma_3)$ allow for balancing the trade-offs between these objectives.

In this section, we introduce the "Multi-Scale Federated DRL Framework," designed to address the challenges posed by stringent latency requirements and limited training data for accurate DRL models. Moreover, it tackles the issues of suboptimal decisions by newly activated V2V pairs and potential obsolescence of well-trained DRL models due to vehicle mobility.

The proposed framework leverages the similarities in channel quality and environmental observations among nearby V2V pairs through a multi-scale approach. It combines centralized clustering on a large timescale with federated DRL on a small scale, aiming to train robust DRL models and enhance the performance of newly activated V2V/V2I pairs. The ultimate goal is to optimize V2X communication in vehicular networks.

We propose a novel multi-scale decentralized federated DRL which synergizes federated learning and DRL techniques to address the mode selection and resource allocation challenges in vehicular networks. Fig. 3 illustrates the architecture of the multi-scale decentralized federated DRL framework, comprising two distinct procedures operating at different scales.

For a centralized training and for less periodic timescale, the base station periodically constructs undirected graphs based on large-scale channel gains and clusters nearby with similar channel conditions. Additionally, each cluster's candidate resource block group is determined to minimize network dimension and mitigate resource conflicts.

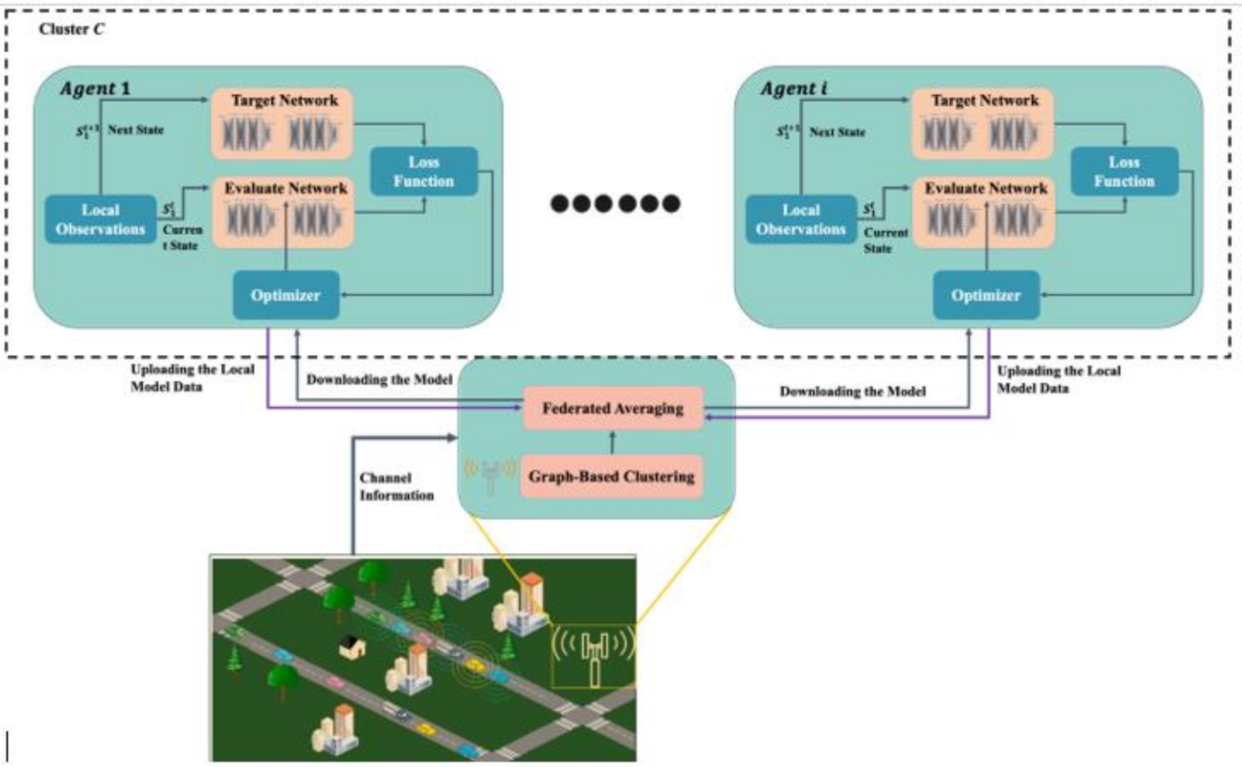

Figure 2.Proposed Federated DRL Architecture.

For local agents, federated learning is employed to collaboratively average local models of V2V and V2I pairs within each cluster. V2V and V2I pairs in the same cluster independently select actions and train their local models in each subframe. Periodically, every few hundreds of subframes, the local models of member pairs within a cluster are uploaded, averaged, and then shared as a global network with all members. Notably, the global network can be efficiently downloaded by newly activated pairs to expedite their deployment without time-consuming training processes.

The procedure for clustering is as follows. Initially, we create an undirected graph $\mathcal{G}(\mathcal{V},\mathcal{E})$, where each V2V/V2I pair represents a vertex, and edges connect pairs of vertices. $\mathcal{V}(\mathcal{G})$ and $\mathcal{E}(\mathcal{G})$ denote the sets of vertices and edges, respectively. Given the unreliable link connections between nearby connections vehicular networks due to blockage, we use large-scale channel gains as edge weights instead of Euclidean distances. The weight of the edge between vertex α and β is defined as [2]:

$$w_{\alpha,\beta} = \max\left(g_{\alpha,\beta}, g_{\beta,\alpha}\right) \quad \text{for all } \beta \neq \alpha. \tag{19}$$

To clusters with similar channel gains, we formulate the clustering problem as a graph partitioning problem aiming to maximize the sum of weights of edges inside clusters. The objective function can be expressed as:

$$\max_{C_1,\ldots,C_C} \sum_{c=1}^{C} \sum_{i,j\in C_c} w_{i,j}$$

subject to the constraints:

$$C_1 \cup C_2 \cup \ldots \cup C_C = \mathcal{V}(\mathcal{G})$$
$$C_i \cap C_j = 0 \quad \text{for all } i \neq j$$

where $\mathcal{V}(\mathcal{G})$ and $\mathcal{E}(\mathcal{G})$ represent the sets of vertices and edges of the undirected graph $\mathcal{G}$, respectively.

The above graph partitioning problem is known to be NP-hard, and traditional Euclidean distance-based clustering methods like K-means and K-medoids are not applicable due to the weights in the constructed undirected graph being based on channel gains instead of Euclidean distances. To address these challenges, we adopt the spectral clustering method, which utilizes similarity-based weights and finds an optimal solution through multiple searches [2][50]. To mitigate interference among V2V and V2I pairs in the same cluster, it is essential to allocate orthogonal resources to them. Based on the clustering results, we define the candidate RB group for cluster C_c as:

$$F_c = \mathcal{F} \setminus \{m \mid m \in \mathcal{M}, m \in C_c\}$$

This approach aims to train robust DRL models in local agents and improve the performance of newly activated V2V or V2I pairs. With the cluster sets and candidate RB groups obtained from the centralized clustering, we introduce federated learning to facilitate the training of robust DRL models. The federated DRL process consists of numerous coordination rounds. During each coordination round $r = 1, 2, \ldots$, the base station BS distributes the pretrained or averaged model to V2V/V2I pairs in the same clusters. Each pair then performs DRL-Based

Decentralized Learning Algorithm to train their own local models based on local training data. After training, the BS selects pairs from the same cluster to upload their local models. Federated averaging is then performed to calculate the weights of the global Q network, which is then later redistributed to all pairs in the cluster.

In the federated DRL process, each pair independently selects its action based on local observations, without any knowledge of actions taken by other pairs. This may limit the ability to characterize the entire environment, potentially leading to resource collisions and suboptimal decisions. To address this, we introduce an asynchronous scheme where subframes are divided into multiple subframe blocks. Each pair in the same cluster is allocated to a specific subframe and asynchronously performs action selection at the designated subframe.

For newly activated pairs, they request the BS to decide the cluster set to which they belong. The global DRL model and detailed network parameters of their specific clusters are then downloaded to these newly activated pairs. By doing so, the time-consuming training process of local DRL models is avoided, and they can quickly integrate into the existing federated DRL framework.

The core process of federated DRL is achieved through minibatch-based stochastic gradient descent for federated averaging. The global model's weights are updated based on the local models' weights from pairs within the same cluster. The update process occurs with a soft update factor τ, which stabilizes the learning process and ensures that the parameters of the target network(φ_t) are slowly updated compared to the evaluate network (φ_e). We can summarize the reward function of local agents as follow:

$$R(s,a) = \lambda_1 \sum_{m\in M} U\left(R_{s,m} - R_{\min,m}\right) + \\ \lambda_2 \sum_{\alpha\in A} U\left(Q_{s,\alpha} - Q_{\max}\right) + \\ \lambda_3 \sum_{\alpha\in A} U\left(\frac{P_{s,\alpha}\times g_{\alpha}[f]}{\sigma^2+\sum_{\beta}\left(\rho_{s,\beta}\times P_{s,\beta}\times g_{\beta}[f]\right)+\sum_{\gamma}\left(\rho_{\gamma,\alpha}\times P_{\gamma,f}\times g_{\gamma,B}[f]\right)} - \\ \frac{\gamma_0}{\ln(1-p_0)}\right)$$

Where:

- s represents the current state.
- a denotes the action taken in state s .
- M is the set of resource blocks RBs.
- A is the set of available actions.
- $R_{s,m}$ is the received signal-to-noise ratio SNR of $RB(m)$ in state s .
- $R_{\min,m}$ is the minimum required SNR of $RB(m)$.
- $Q_{s,\alpha}$ is the channel quality of action α in state s .
- $Q_{\max}$ is the maximum allowable channel quality.
- $P_{s,\alpha}$ is the transmit power for action α in state s.
- $g_{\alpha}[f]$ represents the channel gain of action α at frequency f .
- σ^2 denotes the total interference and noise power.
- $\rho_{s,\beta}$ and $\rho_{\gamma,\alpha}$ are binary variables that indicate if a V2V or V2I belongs to cluster β or γ, respectively.
- $P_{\gamma,f}$ is the transmit power of V2I γ at frequency f .
- $g_{\gamma,B}[f]$ is the channel gain between V2I pairs γ and the BS at frequency f .
- γ_0 is a parameter representing a threshold value.
- p_0 is the target outage probability.
- λ_1, λ_2, and λ_3 are weighting factors for the different components of the reward function.

With this new formula, the reward function captures the trade-offs between the signal quality, power consumption, and resource allocation, helping the federated DRL-based algorithm make more informed decisions during the learning process. By combining the centralized clustering on a large scale with federated DRL on a small-scale and shorter timeframe, the multi-scale decentralized framework, could lead to improved resource allocation and overall network performance.

4. Simulations and Results

In this section, we evaluate the performances of the proposed multi-scale decentralized federated DRL algorithms for cellular vehicle-to-everything (V2X) communications through simulations.

For our simulation study, we chose the SUMO, a widely used open-source microscopic traffic simulator. SUMO allows us to model and simulate vehicular movements and traffic scenarios with high fidelity, making it an ideal choice for evaluating the performance of our proposed multi-scale decentralized federated DRL framework in vehicular networks. To implement our multi-scale decentralized federated DRL framework and interact with SUMO dynamically, we utilized FLOW, a framework that provides deep reinforcement learning-related APIs to work seamlessly with SUMO. FLOW simplifies the integration of reinforcement learning techniques with traffic simulations, enabling us to design and evaluate our DRL-based algorithms efficiently.

To facilitate the development and optimization of our DRL models, we relied on various libraries and tools. Scipy and NumPy provided us with essential functionalities for scientific computing and numerical operations, respectively. Asynchronous RL algorithms allowed us to efficiently train our models by leveraging parallelism and concurrency, speeding up the learning process.

We consider a crossroads scenario in our simulation, where vehicles are distributed based on the spatial Poisson process, and a base station is located at the center of various clusters. Among the vehicles, ten infrastructures simulated transmitting and receiving signals and K active V2V transmitters are randomly selected, and each V2V transmitter establishes a V2V link with the farthest vehicle in its broadcast range. We adopt the large-scale channel gains for the link for V2V pairs, considering their unreliability due to blockage. The communication parameters are based on the urban street scenario in 3GPP TR 37.885. We defined the safety-critical messages of 1060 bytes for latency and reliability requirements of 10ms and 99% with an outage threshold of 5dB. The capacity requirement of V2I defined as 5 bps/Hz. The number of predefined clusters is set as 20. The specific parameters used in the simulations are listed in Table I.

Table 1. Simulation Parameters

Parameter	**Value**
Carrier Freq.	5.9 GHz
# of Channels	15
Channel Bandwidth	1 MHz
# of Resource Blocks	20
# of Clusters	10
# of V2I Pairs	20
# V2V Pairs in each cluster	5 - 45
Path Loss Model	Line of Sight: 44.23 + 16.7log (distance) None-LOS: 42.52 + 30log (distance)
Transmit Power	23 dBm
Noise Power	-114 dBm
Network Update Frequency	2
Federated Averaging Freq.	200
Weights in reward function	0.2, 0.8, 1, 1.2
Discount Factor	0.7
Learning Rate	0.001
Initial and Final Exploration	1, 0.01
Total # of steps	2000

For the simulations, we employ a fully connected neural network as the DRL model [2]. It consists of an input layer, a hidden layer with 256 neurons, and an output layer. ReLU $(f(x) = \max(0,x))$ is used as the activation function, and adaptive moment estimation is the optimizer. The parameters related to the DRL model are provided in Table I. We considered various parameters to evaluate the performance and three algorithms are considered for comparison in this work including a random C-V2X resource selection algorithm [52], a greedy approach where agents always select the channel with the lowest interference, and DRL based algorithm [23]. We evaluate the performance of the proposed algorithms for data rate, and the reliability and latency requirements.

First, we focus on assessing the performance of the proposed algorithm in terms of data rate. In Fig. 4, shows the data rate versus number of created V2V/V2I links and it obvious that the average date rate experiences a decline as the number of communications pairs increases. The increased number of pairs sharing the same channel intensifies interference, thereby leading to reduced transmit rates for the links. However, the proposed method outperforms other approaches in mitigating interference. By employing clustering, coordination among vehicles on different channels and selecting appropriate power levels based on local observations, the proposed method effectively alleviates interference.

On the other hand, the random selection algorithm performs poorly, and the greedy method shows almost equally unsatisfactory results. The random selection fails to consider channel quality and resource allocation, while the greedy method compels V2V pairs to utilize maximum power, further degrading date rates. The strength of the proposed method lies in its ability to leverage clustering and federated techniques. This enables the method to optimize resource allocation and efficiently manage interference.

Figure 4 illustrates the relationship between data rate and the number of established V2V/V2I links. As the number of communication pairs increases, the average data rate exhibits a gradual decline. This phenomenon is attributed to heightened interference as more pairs share the same channel, resulting in reduced transmission rates for individual links.

However, the proposed method demonstrates superior performance in mitigating interference compared to other approaches. By employing clustering, coordinating vehicles on different channels, and strategically adjusting power levels based on local observations, the proposed method effectively alleviates interference and maintains higher data rates.

In contrast, the random selection algorithm and the greedy method exhibit suboptimal results. The random selection algorithm fails to consider channel quality and resource allocation, leading to inefficient resource utilization. The greedy method, by forcing V2V pairs to use maximum power, exacerbates interference and further degrades data rates.

The proposed method's strength lies in its ability to leverage clustering and federated techniques. These approaches enable the method to optimize resource allocation and effectively manage interference, resulting in improved overall system performance.

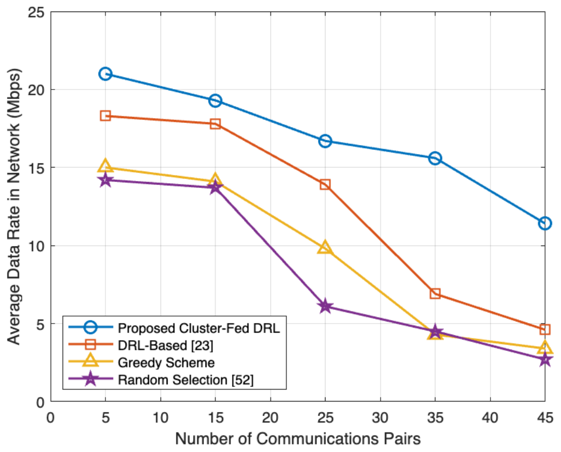

Figure 3. Average Data Rate versus Number of Communications (V2V/V2I) Pairs

We further assess the system's performance by evaluating whether the links meet the latency and quality requirements (success rate). As shown in Figure 5, the proposed algorithm exhibits high performance. The effectiveness of the proposed algorithm lies in its ability to accurately identify unstable links and make optimal transmission mode selections based on local observations and also clustering approach. As a result, the algorithm demonstrates improved performance even as the number of V2V pairs increases.

Moreover, when V2V pairs selects the V2I mode, the algorithm efficiently manages transmit power to meet reliability requirements. This approach effectively reduces interference levels, especially in scenarios with a large number of communication pairs. Consequently, as the number of links grows, the performance gap between the proposed algorithm and other alternatives becomes more apparent. Overall, the simulation results indicate that the proposed algorithm is robust and capable of delivering satisfactory data rates, latency, and quality of service, making it a promising solution for enhancing vehicular communication systems' performance in real-world scenarios.

Figure 5 illustrates the system's performance in terms of meeting latency and quality requirements (success rate). The proposed algorithm consistently demonstrates superior performance in this regard.

The algorithm's effectiveness stems from its ability to accurately identify unstable links and select optimal transmission modes based on local observations and clustering. This approach enables the algorithm to adapt to changing network conditions and mitigate interference effectively, even as the number of V2V pairs grows.

When V2V pairs choose the V2I mode, the algorithm efficiently manages transmit power to ensure reliable communication. This power control strategy helps to reduce interference levels, particularly in scenarios with a high density of communication pairs.

As the number of links increases, the performance gap between the proposed algorithm and other alternatives becomes more pronounced. The simulation results clearly demonstrate the robustness and efficacy of the proposed algorithm in delivering satisfactory data rates, latency, and quality of service. This makes it a promising solution for enhancing the performance of vehicular communication systems in real-world scenarios.

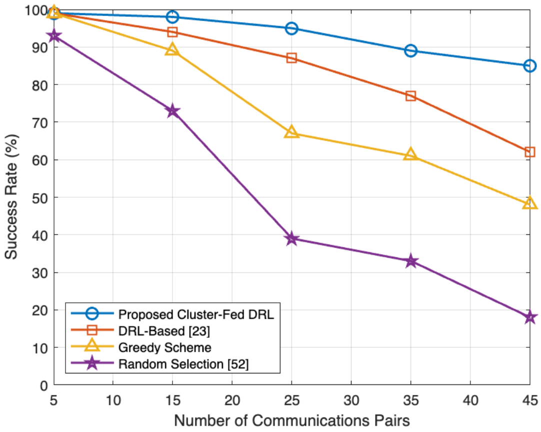

Figure 4. Success Rate versus Number of Communications Pairs

We also conducted evaluations on both data rate and success rate in relation to the SINR threshold. Figures 6 and Fig. 7 illustrates these metrics for varying thresholds, with the number of communications pairs fixed at 25. The proposed algorithm exhibits adaptability to larger thresholds by selecting optimal transmission modes, best reusable channels and adopting appropriate transmission powers, thereby effectively mitigating interference. This adaptability leads to reduced interference and higher reliability, contributing to its overall superior performance. Notably, as the thresholds increase, the average data rate declines. This decrease is attributed to the fact that larger thresholds necessitate communicating pairs to select higher transmission power levels to meet reliability requirements. Consequently, this leads to stronger interference for pairs sharing the same channel. In contrast, the greedy method and random selection schemes always opt for maximum transmission power without considering the transmit rate. As a result, the average rate remains unchanged across different thresholds. This lack of adaptability limits the performance of these two methods.

Figure 6 and Figure 7 depict the relationship between data rate and success rate with varying SINR thresholds, while maintaining a fixed number of communication pairs (25).

The proposed algorithm demonstrates remarkable adaptability to larger SINR thresholds. By strategically selecting optimal transmission modes, identifying the best reusable channels, and adjusting transmission powers, the algorithm effectively mitigates interference. This adaptability results in reduced interference and higher reliability, contributing to its overall superior performance.

However, as the SINR threshold increases, the average data rate gradually declines. This is because higher thresholds necessitate communicating pairs to employ higher transmission power levels to meet reliability requirements. Consequently, this leads to increased interference among pairs sharing the same channel.

In contrast, the greedy method and random selection schemes consistently operate at maximum transmission power, regardless of the SINR threshold. This lack of adaptability limits their performance, as they fail to optimize power usage and mitigate interference effectively. As a result, the average data rate remains relatively unchanged across different thresholds.

Overall, the proposed algorithm's ability to adapt to varying SINR thresholds and optimize transmission parameters is a key factor in its superior performance. This adaptability enables it to achieve higher data rates and reliability, even in challenging communication environments.

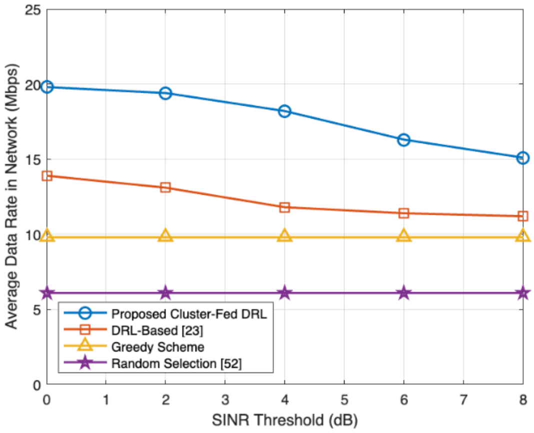

Figure 5. Average Data Rate versus SINR Threshold

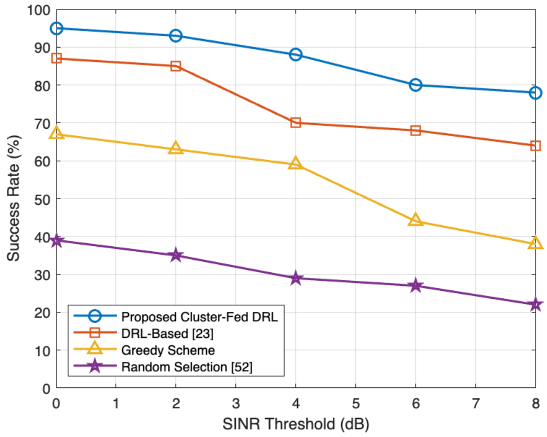

Figure 6. Success Rate versus SINR Threshold

5. Conclusion

In this research, we aimed to empower vehicles with the autonomy to make intelligent channel selection decisions for their transmissions. To achieve this goal, we proposed a novel two-layer multi-agent approach, wherein individual vehicles acted as agents, making autonomous decisions based on their local observations, while clusters of nearby vehicles collaborated to improve overall system performance.

Our decision-making model was built on the foundation of deep reinforcement learning (DRL) techniques, which considered multiple factors such as instantaneous channel state information, queue backlog at the transmitter, interference from different links, and historical selections of nearby links. This allowed the agents to make well-informed choices concerning channel selection and power allocation, optimizing communication efficiency.

To further enhance the performance of our decentralized vehicular agents, we integrated federated learning (FL) into our approach. This facilitated knowledge sharing and synchronization among the individual agents, harnessing the collective intelligence of the network. By consolidating and synchronizing the local models through FL, each agent gained insights into the broader network dynamics beyond its limited observations, leading to more accurate and coordinated decision-making.

The integration of DRL and FL offered several advantages. Firstly, it enabled real-time adaptability for each agent in response to the dynamic nature of the vehicular environment. Secondly, the collective intelligence of the network, harnessed through FL, improved decision-making efficiency, leading to better channel utilization and reduced interference.

The results from our simulations demonstrated the effectiveness of our proposed approach, showcasing its ability to tackle the challenges of vehicular communication systems.

6. Compliance with Ethical Standards

- Authors declare that they have no conflict of interest.
- This article does not contain any studies with human participants or animals performed by any of the authors.

7. Reference:

[1] E. Obiodu, A. Raman, A. K. Abubakar, S. Mangiante, N. Sastry and A. H. Aghvami. (2022, Feb.). DSM-MoC as Baseline: Reliability Assurance via Redundant Cellular Connectivity in Connected Cars. *IEEE Transactions on Network and Service Management*. [Online]. 19(3), pp. 2178-2194. Available: 10.1109/TNSM.2022.3153452

[2] X. Zhang, M. Peng, S. Yan and Y. Sun. (2019, Dec.). Deep-Reinforcement-Learning-Based Mode Selection and Resource Allocation for Cellular V2X Communications. *IEEE Internet of Things Journal*. [Online]. 7(7), pp. 6380-6391. Available: 10.1109/JIOT.2019.2962715

[3] M. Yang, Y. Ju, L. Liu, Q. Pei, K. Yu and J. J. P. C. Rodrigues, “Secure mmWave C-V2X Communications Using Cooperative Jamming," in GLOBECOM 2022 - 2022 IEEE Global Communications Conference, Rio de Janeiro, Brazil, 2022, pp. 2686-2691

[4] T. -X. Zheng *et al*. and K. K. W. (2022, Jun.). Physical-Layer Security of Uplink mmWave Transmissions in Cellular V2X Networks. *IEEE Transactions on Wireless Communications*. [Online]. 21(11), pp. 9818-9833. Available: https://doi.org/10.1109/TWC.2022.3179706

[5] M. S. Bahbahani, E. Alsusa and A. Hammadi. (2022, Nov.). A Directional TDMA Protocol for High Throughput URLLC in mmWave Vehicular Networks. *IEEE Transactions on Vehicular Technology*. [Online]. 72(3), pp. 3584-3599. Available: 10.1109/TVT.2022.3219771

[6] P. Xiang, H. Shan, Z. Su, Z. Zhang, C. Chen and E. -P. Li. (2022, Oct.). Multi-Agent Reinforcement Learning-Based Decentralized Spectrum Access in Vehicular Networks with Emergent Communication. IEEE Communications Letters. [Online]. 27(1), pp. 195-199. Available: https://doi.org/10.1109/LCOMM.2022.3214792

[7] K. Sehla, T. M. T. Nguyen, G. Pujolle and P. B. Velloso. (2022, Mar.). Resource Allocation Modes in C-V2X: From LTE-V2X to 5G-V2X. *IEEE Internet of Things Journal*. [Online]. 9(11), pp. 8291-8314. Available: https://doi.org/10.1109/JIOT.2022.3159591

[8] H. Bagheri et al. and K. Moessner. (2021, Mar.). 5G NR-V2X: Toward Connected and Cooperative Autonomous Driving. IEEE Communications Standards Magazine. [Online]. 5(1), pp. 48-54. Available: https://doi.org/10.1109/MCOMSTD.001.2000069

[9] G. Twardokus and H. Rahbari. (2023, Mar.). Towards Protecting 5G Sidelink Scheduling in C-V2X Against Intelligent DoS Attacks. IEEE Transactions on Wireless Communications. [Online]. 22(11), pp. 7273-7286. Available: https://doi.org/10.1109/TWC.2023.3249665

[10] K. Sehla, T. M. T. Nguyen, G. Pujolle and P. B. Velloso. (2022, Mar.). Resource Allocation Modes in C-V2X: From LTE-V2X to 5G-V2X. IEEE Internet of Things Journal. [Online]. 9(11), pp. 8291-8314.

Available: https://doi.org/10.1109/JIOT.2022.3159591

[11] S. -H. Wu, R. -H. Hwang, C. -Y. Wang and C. -H. Chou. "Deep Reinforcement Learning Based Resource Allocation for 5G V2V Groupcast Communications," in 2023 International Conference on Computing, Networking and Communications (ICNC), Honolulu, HI, USA, 2023, pp. 1-6.

[12] X. Lin, J. G. Andrews, A. Ghosh and R. Ratasuk. (2014, Apr.). An overview of 3GPP device-to-device proximity services. IEEE Communications Magazine. [Online]. 52(4), pp. 40-48. Available: https://doi.org/10.1109/MCOM.2014.6807945

[13] S. Chen et al. and R. Zhao. (2017, Jul.). Vehicle-to-Everything (v2x) Services Supported by LTE-Based Systems and 5G. IEEE Communications Standards Magazine. [Online]. 1(2), pp. 70-76. Available: https://doi.org/10.1109/MCOMSTD.2017.1700015

[14] T. Ran, "Study on Evaluation Methodology of New V2X Use Cases for LTE and NR," Dubrovnik, Croatia, 2017.

[15] M. H. C. Garcia et al. and T. Şahin. (2021, Feb.). A Tutorial on 5G NR V2X Communications. IEEE Communications Surveys & Tutorials. [Online]. 23(3), pp. 1972-2026. Available: https://doi.org/10.1109/COMST.2021.3057017

[16] T. Shahgholi, A. Sheikhahmadi, K. Khamforoosh, and S. Azizi. (2021, Feb.). LPWAN-based hybrid backhaul communication for intelligent transportation systems. Architecture and performance evaluation. EURASIP Journal on Wireless Communications and Networking. *[Online]. (35). Available: https://doi.org/10.1186/s13638-021-01918-2*

[17] H. Yang, L. Zhao, L. Lei and K. Zheng, "A two-stage allocation scheme for delay-sensitive services in dense vehicular networks," IEEE International Conference on Communications Workshops (ICC Workshops), Paris, France, 2017, pp. 1358-1363. Available: https://doi.org/10.1109/ICCW.2017.7962848

[18] L. Liang, S. Xie, G. Y. Li, Z. Ding and X. Yu. (2018, Feb.). Graph-Based Resource Sharing in Vehicular Communication. IEEE Transactions on Wireless Communications. [Online]. 17(7), pp. 4579-4592. Available: https://doi.org/10.1109/TWC.2018.2827958

[19] C. Chen, B. Wang and R. Zhang. (2018, Oct.). Interference Hypergraph-Based Resource Allocation (IHG-RA) for NOMA-Integrated V2X Networks. IEEE Internet of Things Journal. [Online]. 6(1), pp. 161-170. Available: https://doi.org/10.1109/JIOT.2018.2875670

[20] B. Bai, W. Chen, K. B. Letaief and Z. Cao. (2010, Dec.). Low Complexity Outage Optimal Distributed Channel Allocation for Vehicle-to-Vehicle Communications. IEEE Journal on Selected Areas in Communications. [Online]. 29(1), pp. 161-172. Available: https://doi.org/10.1109/JSAC.2011.110116

[21] L. Liang, J. Kim, S. C. Jha, K. Sivanesan and G. Y. Li. (2017, May.). Spectrum and Power Allocation for Vehicular Communications With Delayed CSI Feedback. IEEE Wireless Communications Letters. [Online]. 6(4), pp. 458-461. Available: https://doi.org/10.1109/LWC.2017.2702747

[22] H. Ye, G. Y. Li and B. -H. F. Juang. (2019, Feb.). Deep Reinforcement Learning Based Resource Allocation for V2V Communications. IEEE Transactions on Vehicular Technology. [Online]. 68(4), pp. 3163-3173. Available: https://doi.org/10.1109/TVT.2019.2897134

[23] Y. Wang, X. Zheng and X. Hou, "A Novel Semi-Distributed Transmission Paradigm for NR V2X," in IEEE Globecom Workshops (GC Wkshps), Waikoloa, HI, USA, 2019, pp. 1-6.

[24] H. Mosavat-Jahromi, Y. Li, L. Cai and L. Lu, "NC-MAC: Network Coding-based Distributed MAC Protocol for Reliable Beacon Broadcasting in V2X," GLOBECOM 2020 - 2020 IEEE Global Communications Conference, Taipei, Taiwan, 2020, pp. 1-6.

[25] A. Zappone, M. Di Renzo and M. Debbah. (2019, Jun.). Wireless Networks Design in the Era of Deep Learning: Model-Based, AI-Based, or Both?. IEEE Transactions on Communications. [Online]. 67(10), pp. 7331-7376. Available: https://doi.org/10.1109/TCOMM.2019.2924010

[26] L. Wang, H. Ye, L. Liang and G. Y. Li. (2020, Mar.). Learn to Compress CSI and Allocate Resources in Vehicular Networks. IEEE Transactions on Communications. [Online]. 68(6), pp. 3640-3653. Available: https://doi.org/10.1109/TCOMM.2020.2979124

[27] Lu, Yan, Ping Wang, Shuai Wang, and Wangding Yao, "A Q-learning based SPS resource scheduling algorithm for reliable C-V2X communication," in *5th International Conference on Digital Signal Processing*, 2021, pp. 201-206.

[28] L. Busoniu, R. Babuska and B. De Schutter. (2008, Feb.). A Comprehensive Survey of Multiagent Reinforcement Learning. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews). [Online]. 38(2), pp. 156-172. Available: https://doi.org/10.1109/TSMCC.2007.913919

[29] B. Gu, W. Chen, M. Alazab, X. Tan and M. Guizani. (2022, Jul.). Multiagent Reinforcement Learning-Based Semi-Persistent Scheduling Scheme in C-V2X Mode 4. IEEE Transactions on Vehicular Technology. [Online]. 71(11), pp. 12044-12056. Available: https://doi.org/10.1109/TVT.2022.3189019

[30] L. Cao and H. Yin, "Resource Allocation for Vehicle Platooning in 5G NR-V2X via Deep Reinforcement Learning," in 2021 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom), Bucharest, Romania, 2021, pp. 1-7.

[31] Y. Yuan, G. Zheng, K. -K. Wong and K. B. Letaief. (2021, Jul.). Meta-Reinforcement Learning Based Resource Allocation for Dynamic V2X Communications. IEEE Transactions on Vehicular Technology. [Online]. 70(9), pp. 8964-8977. Available: https://doi.org/10.1109/TVT.2021.3098854

[32] Z. He, L. Wang, H. Ye, G. Y. Li and B. -H. F. Juang, "Resource Allocation based on Graph Neural Networks in Vehicular Communications," GLOBECOM 2020 - 2020 IEEE Global

Communications Conference, Taipei, Taiwan, 2020, pp. 1-5.

[33] R. Wang, X. Jiang, Y. Zhou, Z. Li, D. Wu, T. Tang, A. Fedotov and V. Badenko. (2022, Jun.). Multi-agent reinforcement learning for edge information sharing in vehicular networks. Digital Communications and Networks. [Online]. 8(3), pp. 267-277. Available: https://doi.org/10.1016/j.dcan.2021.08.006

[34] C. Wu, Z. Liu, F. Liu, T. Yoshinaga, Y. Ji and J. Li. (2020, Jun.). Collaborative Learning of Communication Routes in Edge-Enabled Multi-Access Vehicular Environment. IEEE Transactions on Cognitive Communications and Networking. [Online]. 6(4), pp. 1155-1165. Available: https://doi.org/10.1109/TCCN.2020.3002253

[35] B. Gu, X. Yang, Z. Lin, W. Hu, M. Alazab and R. Kharel. (2020, Sep.). Multiagent Actor-Critic Network-Based Incentive Mechanism for Mobile Crowdsensing in Industrial Systems. IEEE Transactions on Industrial Informatics. [Online]. 17(9), pp. 6182-6191. Available: https://doi.org/10.1109/TII.2020.3024611

[36] L. Liang, H. Ye and G. Y. Li. (2019, Aug.). Spectrum Sharing in Vehicular Networks Based on Multi-Agent Reinforcement Learning. IEEE Journal on Selected Areas in Communications. [Online]. 37(10), pp. 2282-2292. Available: https://doi.org/10.1109/JSAC.2019.2933962

[37] J. Tian, Q. Liu, H. Zhang and D. Wu. (2021, Jun.). Multiagent Deep-Reinforcement-Learning-Based Resource Allocation for Heterogeneous QoS Guarantees for Vehicular Networks. IEEE Internet of Things Journal. [Online]. 9(3), pp. 1683-1695. Available: https://doi.org/10.1109/JIOT.2021.3089823

[38] X. Chen et al. and Y. Zhang. (2020, Jan.). Age of Information Aware Radio Resource Management in Vehicular Networks: A Proactive Deep Reinforcement Learning Perspective. IEEE Transactions on Wireless Communications. [Online]. 19(4), pp. 2268-2281. Available: https://doi.org/10.1109/TWC.2019.2963667

[39] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Arcas, “Communication-efficient learning of deep networks from decentralized data,” in *Artificial intelligence and statistics, 2017*, pp. 1273-1282.

[40] S. Niknam, H. S. Dhillon and J. H. Reed. (2020, Jun.). Federated Learning for Wireless Communications: Motivation, Opportunities, and Challenges. IEEE Communications Magazine. [Online]. 58(6), pp. 46-51. Available: https://doi.org/10.1109/MCOM.001.1900461

[41] J. Qi, Q. Zhou, L. Lei and Kan Zheng, “Federated reinforcement learning: Techniques, applications, and open challenges,” 2021.

[42] R. Valente, C. Senna, P. Rito and S. Sargento. (2023, Feb.). Embedded Federated Learning for VANET Environments. *Applied Sciences.* [Online]. 13(4), p. 2329. Available: https://doi.org/10.3390/app13042329

[43] G. Wang, X. Fangmin, H. Zhang and C. Zhao. (2022, Apr.). Joint resource management for mobility supported federated learning in Internet of Vehicles. *Future Generation Computer Systems.* [Online]. 129, pp. 199-211. Available: https://doi.org/10.1016/j.future.2021.11.020

[44] Q. Guo, F. Tang and N. Kato. (2022, Sep.). Federated Reinforcement Learning-Based Resource Allocation in D2D-Enabled 6G. IEEE Network. [Online]. 37(5), pp. 89-95. Available: https://doi.org/10.1109/MNET.122.2200102

[45] S. Kandukuri and S. Boyd. (2002, Jan.). Optimal power control in interference-limited fading wireless channels with outage-probability specifications. IEEE Transactions on Wireless Communications. [Online]. 1(1), pp. 46-55. Available: https://doi.org/10.1109/7693.975444

[46] X. Li, L. Lu, W. Ni, A. Jamalipour, D. Zhang and H. Du. (2022, May.). Federated Multi-Agent Deep Reinforcement Learning for Resource Allocation of Vehicle-to-Vehicle Communications. IEEE Transactions on Vehicular Technology. [Online]. 71(8), pp. 8810-8824. Available: https://doi.org/10.1109/TVT.2022.3173057

[47] J. M. Meredith, “Technical Specification Group Radio Access Network; Study on LTEBased V2X Services; (Release 14),” 3rd Generation Partnership Project, 2016.

[48] C. Guo, L. Liang and G. Y. Li. (2019, Feb.). Resource Allocation for Low-Latency Vehicular Communications: An Effective Capacity Perspective. IEEE Journal on Selected Areas in Communications. [Online]. 37(4), pp. 905-917. Available: https://doi.org/10.1109/JSAC.2019.2898743

[49] J. Zeng et al and Y. Wu, “Multi-D3QN: A multi-strategy deep reinforcement learning for service composition in cloud manufacturing,” *International Conference on Collaborative Computing: Networking, Applications and Worksharing*. Cham: Springer International Publishing, 2021, pp. 225-240.

[50] H. Yang, X. Xie and M. Kadoch. (2019, Jan.). Intelligent Resource Management Based on Reinforcement Learning for Ultra-Reliable and Low-Latency IoV Communication Networks. IEEE Transactions on Vehicular Technology. [Online]. 68(5), pp. 4157-4169. Available: https://doi.org/10.1109/TVT.2018.2890686

[51] S. Chen, J. Hu, Y. Shi and L. Zhao. (2016, Sep.). LTE-V: A TD-LTE-Based V2X Solution for Future Vehicular Network. IEEE Internet of Things Journal. [Online]. 3(6), pp. 997-1005. Available: https://doi.org/10.1109/JIOT.2016.2611605

[52] L. Liang, S. Xie, G. Y. Li, Z. Ding and X. Yu. (2018, Apr.). Graph-Based Resource Sharing in Vehicular Communication. IEEE Transactions on Wireless Communications. [Online]. 17(7), pp. 4579-4592. Available: https://doi.org/10.1109/TWC.2018.2827958